Importance Sampling of Probabilistic Contracts in Web Services

Ajay Kattepur

IRISA/INRIA, Campus Universitaire de Beaulieu, Rennes, France.

Abstract. With web services quality of service (QoS) modeled as random variables, the accuracy of sampled values for precise service level agreements (SLAs) come into question. Samples with lower spread are more accurate for calculating contractual obligations, which is typically not the case for web services QoS. Moreover, the extreme values in case of heavy-tailed distributions (eg. 99.99 percentile) are seldom observed through limited sampling schemes. To improve the accuracy of contracts, we propose the use of variance reduction techniques such as importance sampling. We demonstrate this for contracts involving *demand* and *refuel* operations within the *Dell* supply chain example. Using measured values, efficient forecasting of future deviation of contracts may also be performed. A consequence of this is a more precise definition of sampling, measurement and variance tolerance in SLA declarations.

Keywords: Web Services, QoS, Importance Sampling, SLA.

1 Introduction

Web services continue to attract applications in many areas [1]. With increasing efforts to standardize performance of web services, focus has shifted to Quality of Service (QoS) levels. This is important to consider in case of orchestrations that specify the control flow for multiple services. To this end, contractual guarantees and service level agreements (SLAs) [2] are critical to ensure adequate QoS performance.

QoS metrics being random variables, the treatment of contractual obligations tends toward probabilistic criterion [3]. Contractual obligations may be specified as varying percentile values of such distributions rather than "hard" values. In [4], composition and monitoring such contracts with stochastic dominance have been examined.

As metrics such as response time and throughput rates can have heavy tails, estimating extreme values becomes difficult with few observations. The *availability* of a web service might need contracts for extreme percentiles in the response time profile (99.99 percentile). For instance, an ambulance or disaster management web service must be available 24×7, indicating a high availability requirement. These values are dependent on sampled random values and can lead to high variance in contractual guarantees.

The use of *importance sampling* [5] is proposed as a solution to these problems. Disadvantages of conventional Monte-Carlo techniques such as high variance of percentile values may be eliminated. In case of heavy tailed distributions, unobserved extreme percentiles can be quantified with higher accuracy. These are stochastically "important" observations to estimate contractual deviations. These issues are demonstrated with the *Dell* example [6], a choreography involving *Dell Plant* and *Supplier* orchestrations. We study more accurate bounds for supplier contracts with varying plant demand rates. Further, we show how QoS metrics such as stock level deviations (specially long delays) can be estimated with low variance.

The rest of the paper is organized as follows: Section 2 introduces the probabilistic contract composition procedure for web services' QoS. Importance sampling is briefly introduced in Section 3 with emphasis on contractual sampling in web services and sample deviations. The Dell application is introduced in Section 4 with two workflows interacting in a choreography. The two application of importance sampling with respect to the Dell supply chain are described in Sections 4.1 and 4.2. Related work and conclusions of the paper are included in Sections 5 and 6, respectively.

2 Probabilistic QoS Contracts

Available literature on industry standards in QoS [7] provides a family of QoS metrics that are needed to specify SLAs. These can be subsumed into the following four general QoS observations: *Service Latency*, *Per Invocation Cost*, *Output Data Quality* and *Inter-Query Intervals*. To handle such diverse domains, metrics and algebra for QoS, a framework is proposed in [4]. Using such an algebra, QoS metrics may be defined explicitly with domains, increments and comparisons within service orchestrations.

For a domain $\mathbb{D}_Q$ of a QoS parameter Q, behavior can be represented by its distribution F_Q:

$$F_Q(x) = \mathbb{P}(Q \leq x) \tag{1}$$

Making use of stochastic ordering [8], this is refined for probability distributions F and G over a totally ordered domain $\mathbb{D}$:

$$G_Q \preceq F_Q \quad \iff \quad \forall x \in \mathbb{D}_Q, \quad G_Q(x) \geq F_Q(x) \tag{2}$$

That is, there are more chances of being less than x (partial order $\preceq$) if the random variable is drawn according to G than according to F. A QoS contract must specify the obligations of the two parties:

- The obligations that the orchestration has regarding the service are seen as *assumptions* by the service - the orchestration is supposed to meet them.
- The obligations that the service has regarding the orchestration are seen as *guarantees* by the service - the service commits to meeting them as long as assumptions are met.

Definition 1 *A probabilistic contract is a pair (Assumptions, Guarantees), which both are lists of tuples $(Q, \mathbb{D}_Q, F_Q)$, where Q is a QoS parameter with QoS domain $\mathbb{D}_Q$ and distribution F_Q.*

Once contracts have been agreed, they must be monitored by the orchestration for possible violation as described in [3].

3 Importance Sampling

In case of web services' SLAs, these rare event simulations can be used to determine the occurrence of failure or deviation from contracts. Traditional Monte-Carlo (MC) methods waste a lot of time in a region of the state space which is "far" from the rare set of interest. Modifying the underlying distributions to move "near" the states of interest provides a more efficient means of analysis. With typical Monte-Carlo (MC), if the mean $\mu = 10^{-5}$ and if we want the expected number of occurrences of this event to be at least 100, we must take approximately $N = 10^7$ runs. For lower values of N, not even a single occurrence of this event may be seen - leading to the faulty conclusion that the event does not occur.

Importance sampling (IS) [5] increases the probability of the rare event while multiplying the estimator by an appropriate likelihood ratio so that it remains unbiased. Consider the case of a random variable Q with probability density function (PDF) F_Q for which the probability of a rare event $\mathbb{P}(H(Q) > \Phi)$ is to be estimated. Here $H(Q)$ is a continuous scalar function and Φ is the threshold. Using Monte-Carlo, one generates independent and identically distributed samples $Q_1, Q_2, ...Q_N$ from the PDF F_Q and then estimates the probability:

$$\mathbb{P}_{MC} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{H(Q_i)>\Phi} \tag{3}$$

where $\mathbf{1}_{H(Q)>\Phi}$ is 1 if $H(Q) > \Phi$ and 0 otherwise. For a rare event, such a technique needs many runs for low variance estimates.

With Importance Sampling (IS) [5], variance can be reduced without increasing the number of samples. The idea is to generate samples $Q_1, Q_2, ...Q_N$ from an auxiliary PDF G_Q and then estimate probability:

$$\mathbb{P}_{IS} = \frac{1}{N} \sum_{i=1}^{N} H(Q_i) \mathbf{1}_{H(Q_i)>\Phi} \frac{F_Q(Q_i)}{G_Q(Q_i)} \tag{4}$$

It is evident that G_Q should be chosen such that it has a thicker tail than F_Q. If F_Q is large over a set but G_Q is small, then $\left(\frac{F_Q}{G_Q}\right)$ would be large and it would result in a large variance. It is useful if we can choose G_Q to be similar to F_Q in terms of shape. Analytically, we can show that the best G_Q is the one that would result in a variance that is minimized [5]. In order to perform this selection, some

sort of knowledge about the distribution is assumed, either through theory or pre-collected statistical data.

As in the case of most statistical techniques, the monitoring of contracts is also based on *samples* of the *population* of QoS. If the variance in values of the sample set is large then the mean is not as representative of the data as if the spread of data is small. If only a sample is given and we wish to make a statement about the population standard deviation(from which the sample is drawn), then we need to use the sample standard deviation. If $Q_1, Q_2, ..., Q_N$ is a sample of N observations, the sample variance is given by:

$$s^2 = \frac{\sum_{i=1}^{N}(Q_i - \bar{Q}^2)}{N-1} \tag{5}$$

with $\bar{Q}$ as the sample mean. This sample standard deviation can be used to represent the deviation in the population QoS output and is used in this paper.

4 Dell Supply Chain

To demonstrate the variation in the QoS domains in real-world services, we study the *Dell* example [6]. The *Dell* application is a system that processes orders from customers interacting with the Dell webstore. According to [6], this consists of the following prominent entities:

- *Dell Plant* - Receive the orders from the Dell webstore and are responsible for the assembly of the components. For this they interact with the *Revolvers* to procure the required items.
- *Revolvers* - Warehouses belonging to Dell which are stocked by the suppliers. Though Dell owns the revolvers, the inventory is owned and managed by the *Suppliers* to meet the demands of the *Dell Plant*.
- *Suppliers* - They produce the components that are sent to the revolvers at Dell. Periodic polling of the *Revolvers* ensure estimates of inventory levels and their decrements.

Essentially, there are a *Dell Plant* and *Supplier* orchestrations that are choreographed through common *Revolvers*. The critical aspect in the Dell choreography is efficient management of revolver levels. It is a shared *buffer* resource that is accessed by both the Dell Plant and the Suppliers. As discussed in [6], for the efficient working of the supply chain, the interaction between the Dell Plant and the Supply-side workflows should be taken into account.

The requests made by the plant for certain items will be favorably replied to if the revolvers have enough stock. This stocking of the revolvers is done independently by the suppliers. The suppliers periodically poll (withdraw inventory levels) from the revolvers to estimate the stock level. In such a case, a contract can be made on the levels of stock that must be maintained in the revolver. The customer side agreement limits the throughput rate. The supplier side agreement ensures constant refueling of inventory levels, which in turn ensures that

the delay time for the customer is minimized. Thus, it represents a *choreography* comprising two plant-side and supplier-side orchestrations interacting via the revolver as a shared resource.

4.1 Contract Composition

For the *Dell* example, as QoS metrics are inherent to the functionality of the choreography, specifying explicitly probabilities of outage is necessary. Proposed are the following two concrete metrics that qualitatively evaluate these workflows:

- **Assumption**: The *demand* (number of orders/hour) distributions from the Dell plant made to a particular revolver. It is the prerogative of the plant to maintain demand within acceptable range of the contracts.
- **Guarantee**: The *delay* (hours) distribution in obtaining products from revolvers. This, in turn, is dependent on the availability of products in the revolver. The suppliers ensure efficient and timely refueling to maintain acceptable delays in the supply chain.

Consider the *assumption* on the query rate of the customer shown as an exponential distribution as in Fig. 1. Repeatedly pinging the service in order to receive boundary values of the distribution is expensive and not reflective of run-time performance. This is demonstrated for three values in Table 1 with 10000 runs. Conventional Monte-Carlo does not detect the probability of inter-query periods being less than 100, 50 or 20 minutes (which can be fallaciously interpreted as the rare event never occurring). Using an importance sampling distribution, accurate mean and sampling variance values are produced for the probability of crossing these thresholds. Such a level of accuracy is needed specially for critical web services (crisis management such as ambulance or fire stations). For conventional web services contracts as well, such precise contractual obligations can reduce the need for extended monitoring of services contracts.

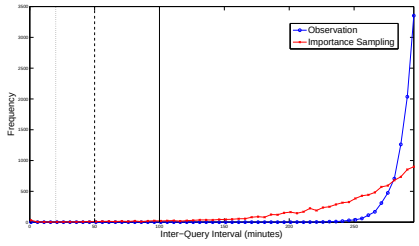

Fig. 1. Inter-query period fitting.

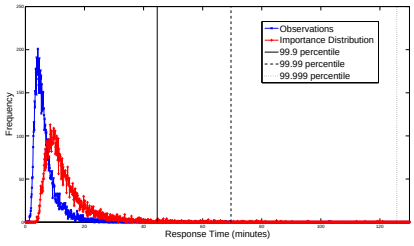

Fig. 2. Response time fitting.

A corresponding *guarantee* from the service provider regarding the response time may be estimated as a long tailed distribution (Fig. 2, Table 2). Once again we concentrate on the outlying percentile values. The outputs for the traditional Monte-Carlo runs produce higher sample variance compared to the importance sampling scheme.

Inter-query period (mins.)	mean MC	variance MC	mean IS	variance IS
100	0	0	0.0086	0.0094
50	0	0	0.0018	7.36×10^{-5}
20	0	0	7.99×10^{-5}	7.37×10^{-6}

Table 1. Inter-query periods by Monte-Carlo (MC) and Importance Sampling (IS).

Percentile	Latency (mins.)	mean MC	variance MC	mean IS	variance IS
99.9	44.61	0.0022	2.456×10^{-6}	0.0018	3.5548×10^{-7}
99.99	69.58	5.2×10^{-4}	5.65×10^{-7}	3.04×10^{-4}	3.82×10^{-8}
99.999	125.70	1.1×10^{-4}	1.19×10^{-7}	3.47×10^{-7}	3.12×10^{-9}

Table 2. Latency by Monte-Carlo (MC) and Importance Sampling (IS) schemes.

An advantage of this scheme is that the contracts will be formulated as explicit probabilities of contractual deviation. The WSLA framework [10], refined with precise probabilistic percentile values of QoS distributions specified as:

```
<Assumptions>
<SLAParameter name="InterQueryPeriod" type="float" unit="seconds" />
<Predicate xsi:type="wsla:Greater">
   <Percentile> 95 </Percentile>   <Value> 30 </Value>
   <SampleVariance> 10^-3 </SampleVariance>  </Predicate>  </Assumptions>
<Guarantees>
<SLAParameter name="ResponseTime" type="float" unit="seconds" />
<Predicate xsi:type="wsla:Less">
   <Percentile> 99 </Percentile>   <Value> 15 </Value>
   <SampleVariance> 10^-3 </SampleVariance>  </Predicate>  </Guarantees>
```

The contract now specifies the contract from the *assumption-guarantee* viewpoint. For any measurement period, the `Percentile` values of the `ResponseTime` should be less than the specified bounds. On the other hand, the `InterQueryPeriod` should be greater than the threshold values. In both cases, the `SampleVariance` is taken into account. Such a framework allows for distributions to be used for both contractual specification and monitoring deviations.

4.2 Forecasting

Traditional forecasting models like autoregressive moving averages [9] rely heavily on accurate mathematical modeling of workflow processes. In this section, we propose using pre-identified contracts / observations to provide an easier method of forecasting outages in web services orchestrations. Consider a Dell revolver with critical stock of 10 items, refueling batch 50 items and a polling period of 10 hours. With an *assumption* distribution of orders/hour shown in Fig. 3, the response time distribution obtained over a period of 1 week is shown in Fig. 4. If an item is available, it is procured immediately. Else, it is refueled with a supplier delay when polling detects sub-critical revolver levels.

In order to develop a *guarantee* distribution, the Dell plant must estimate the probability that delays over 72, 96 or 120 hours are experienced (leading to cancellation in orders). Through importance sampling, these values can be better estimated as in Table 3. Notice that the variance through importance sampling is several orders of magnitude lower than conventional Monte-Carlo. The Dell plant can provision more stringent supplier obligations to reduce the delays. For instance, changing the critical stock to 50 items, refueling batch 200

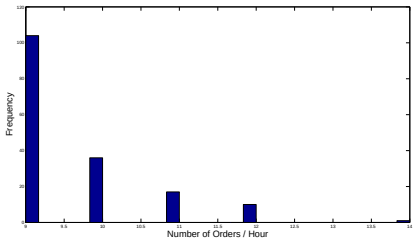

Fig. 3. Assumption: *Plant* side demand distributions.

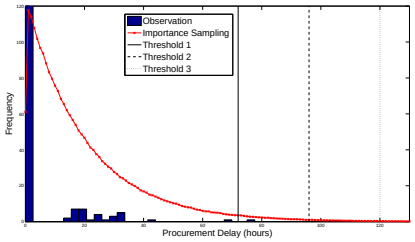

Fig. 4. Guarantee: *Supplier* side procurement delays.

items produces a new set of values, with lower probabilities of crossing outlying values as shown in Table 4.

Such changes produced by improved supplier performance is barely observed through traditional Monte-Carlo sampling, thus proving the efficacy of Importance Sampling. Application of forecasting through pre-negotiated contracts emphasize the need for precise contractual obligations needed in web services.

Delay (hours)	mean MC	variance MC	mean IS	variance IS
72	0.002	3.2×10^{-3}	0.0016	1.72×10^{-7}
96	0	0	3.88×10^{-4}	3.71×10^{-8}
120	0	0	1.02×10^{-4}	6.25×10^{-9}

Table 3. Original contract estimates.

Delay (hours)	mean MC	variance MC	mean IS	variance IS
72	0	0	4.08×10^{-4}	1.35×10^{-8}
96	0	0	9.91×10^{-5}	1.89×10^{-9}
120	0	0	2.734×10^{-5}	6.74×10^{-10}

Table 4. Reformulated contract estimates providing lower probabilities of delay.

5 Related Work

The use of probabilistic QoS and contracts was introduced by Rosario et al [3] and Bistarelli et al [11]. Instead of using hard bound values for parameters such as response time, the authors proposed a probabilistic contract monitoring approach to model the QoS bounds. The composite service QoS was modeled using probabilistic processes by Hwang et al [12] where the authors combine orchestration constructs to derive global probability distributions.

In [14], Gallotti et al propose using a probabilistic model checker to assess non-functional quality attributes of workflows such as performance and reliability. Validating SLA conformance is studied by Boschi et al [15]. A series of experiments to evaluate different sampling techniques in an online environment is studied.

The use of importance sampling to change probability of occurrence of events in well known [5]. An associated work in this area is importance splitting [13]. Importance splitting considers the estimation of a rare event by deploying several conditional probabilities during simulation runs, reducing the need to identify importance distributions as used in this case.

6 Conclusion

QoS aspects are critical to the functioning of most web service orchestrations and choreographies, needing more precise specifications of SLAs. This is difficult as distributions of QoS values have high variance when sampled with inefficient Monte-Carlo techniques. In most cases, the tails of QoS distributions are either neglected or averaged out in contractual specifications. Applying importance sampling to such distributions can provide better estimates of outlying values with relatively low variance. As demonstrated in this paper on the *Dell* supply chain application, importance sampling can have significant imperatives for both contract composition as well as forecasting deviations for critical services. The extension of this approach in case of WSLA specifications are also provided with a precise definition of sample variance.

References

1. G. Alonso, F. Casati, H. Kuno and V. Machiraju, "Web Services: Concepts, Architectures and Applications," *Springer*, 2004.
2. P. Bhoj, S. Singhal and S. Chutani, "SLA management in federated environments," *Symp. on Dist. Mgmt. for the Networked Millennium*, pp. 293–308, 1999.
3. S. Rosario, A. Benveniste, S. Haar and C. Jard, "Probabilistic QoS and soft contracts for transaction based Web services," *IEEE ICWS*, 2007.
4. S. Rosario, A. Benveniste and C. Jard, "Flexible Probabilistic QoS Management of Orchestrations," *Int. J. Web Service Res.*, vol. 7, no. 2, pp. 21–42, 2010.
5. J. A. Bucklew, "Introduction to rare event simulation," *Springer-Verlag*, 2004.
6. R. Kapunscinski, R. Q. Zhang, P. Carbonneau, R. Moore and B. Reeves, "Inventory Decisions in Dells Supply Chain," *Interfaces*, vol. 34, no. 3, pp. 191–205, 2004.
7. World Wide Web Consortium, "QoS for Web Services: Requirements and Possible Approaches," *W3C Working Group Note*, Nov. 2003.
8. M. Shaked and J. G. Shanthikumar, "Stochastic Orders," *Springer Statistics*, 2006.
9. S. Makridakis and S. Wheelwright, "Adaptive Filtering: An Integrated Autoregressive/Moving Average Filter for Time Series Forecasting," *Operational Research Quarterly*, vol. 28, no. 2, pp. 425–437, 1977.
10. H. Ludwig, A. Keller, A. Dan, R. P. King and R. Franck, "Web Service Level Agreement (WSLA) Language Specification," *IBM Corporation*, 2003.
11. S. Bistarelli and F. S. Santini, "Soft Constraints for Quality Aspects in Service Oriented Architectures," *Workshop on Service Oriented Computing*, Italy, 2009.
12. S. Y. Hwang, H. Wang, J. Tang and J. Srivastava, "A probabilistic approach to modeling and estimating the QoS of web-services-based workflows," *Elsevier Information Sciences*, vol. 177, pp. 5484–5503, 2007.
13. J. Morio, R. Pastel and F. Le Gland, "An overview of importance splitting for rare event simulation," *European J. of Physics*, vol. 31, pp. 1295-1303, 2010.
14. S. Gallotti, C. Ghezzi, R. Mirandola and G. Tamburrelli, "Quality Prediction of Service Compositions through Probabilistic Model Checking," *Quality of Software Architectures*, LNCS vol. 5281, pp. 119-134, 2008.
15. E. Boschi, S. Denazis and T. Zseby, "A measurement framework for inter-domain SLA validation," *Elsevier Computer Communications*, vol. 29, pp. 703-716, 2006.